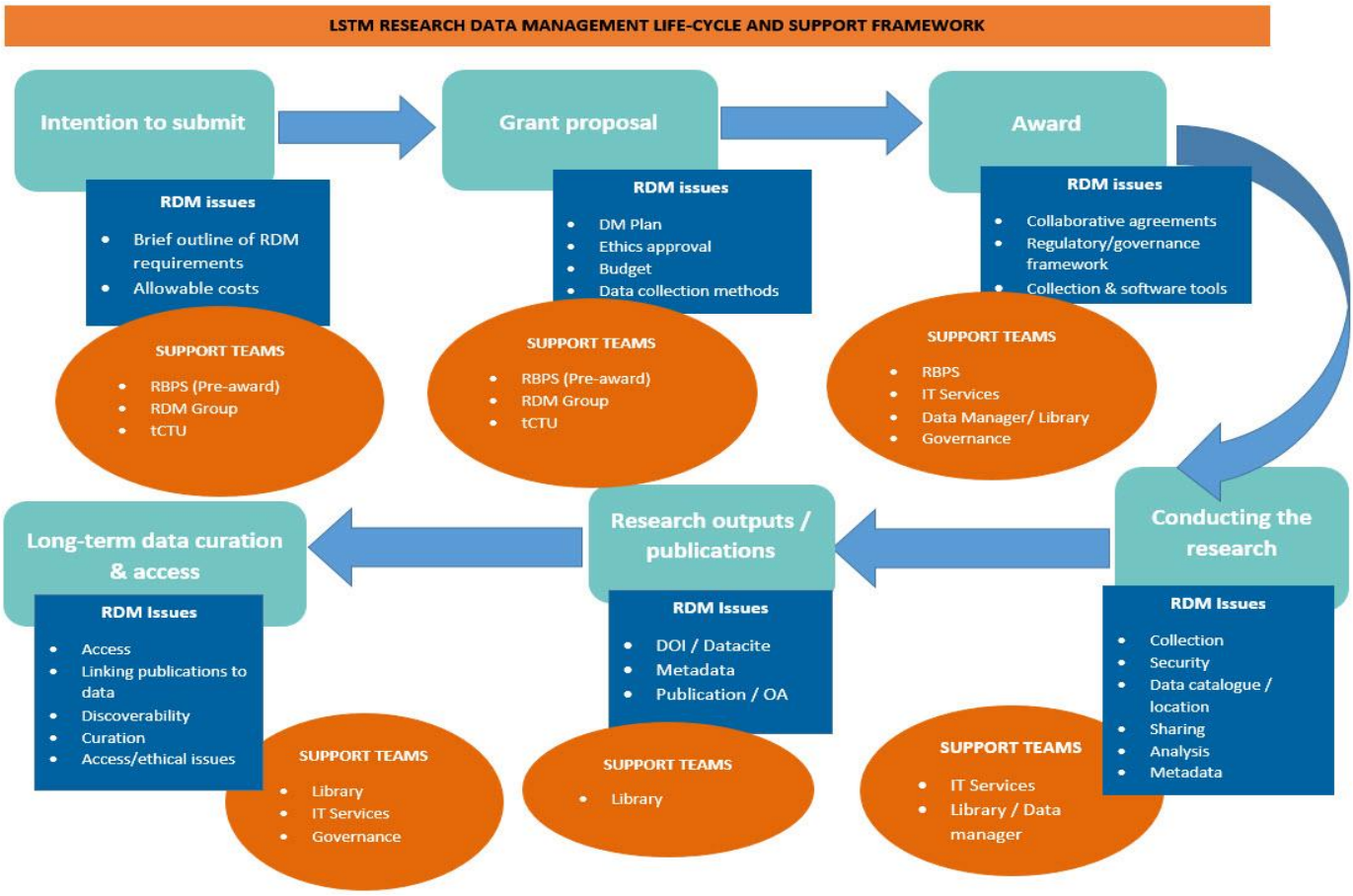
§1 Managing your Data

When beginning a new research project, you will need to consider a wide range of issues relating to the management of research data. Many of these issues will be relevant whether or not the research is funded by an external sponsor. Using a checklist and/or writing a data management plan is an essential first step to take at the pre-project stage. You can then continuously refer to it throughout the project. A useful starting point is the DCC (Digital Curation Centre) Checklist for a Data Management Plan. It presents the main themes and questions that researchers may want to cover when writing a data management plan.

1.1 LSTM RDM lifecycle and support framework

Below is a worflow which helps to illustrate the stages of the grant process from "Intention to submit" to the data outputs which may need long-term curation and access in order to ensure compliance with your funder's policies. this will help you to understand the different teams that may need to assist you during those stages at LSTM.



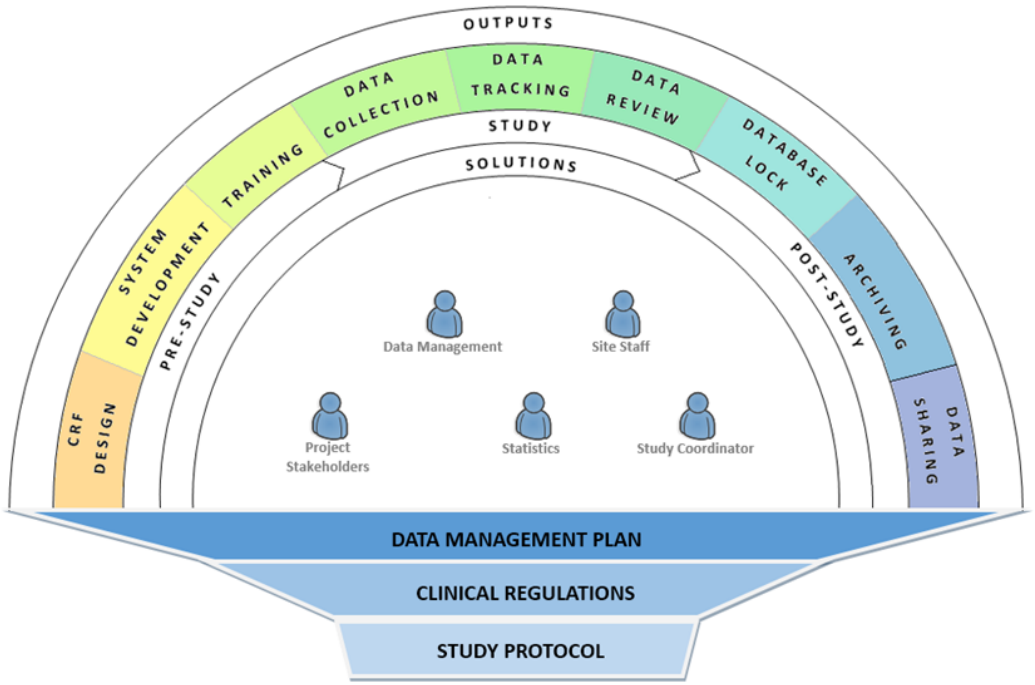
1.2 Other Lifecycles for data management

Although the above is to act as a guide for the entire process, it may be that you need specific guidance during the stage of 'conducting the research'. There are other lifecylces that can be

utilised during that stage such as The Association of Data Management in the Tropics lifecylce which includes Standard Operating Procedures (SOPs).

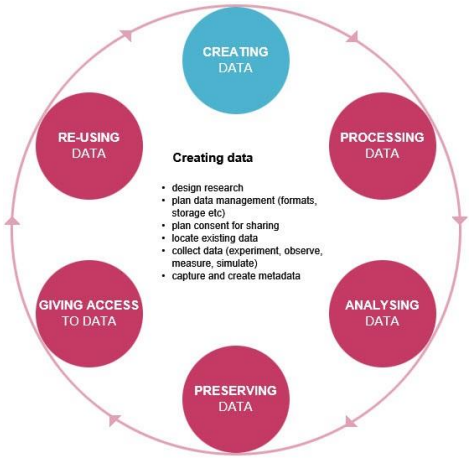
1.2.1 Association of Data Management in the Tropics (ADMIT)

Developed by a collaboration of data management professionals, with a wealth of experience of working in Low and Middle Income (LMIC) settings as part of the Association of Data Management in the Tropics (ADMIT), the lifecycle provides standard operating procedure (SOP) templates for each stage within the lifecycle. The data management plan coupled with the SOPs will form the constraints within which the data management should be carried out.



1.2.2 UK Data Archive Data Lifecylce

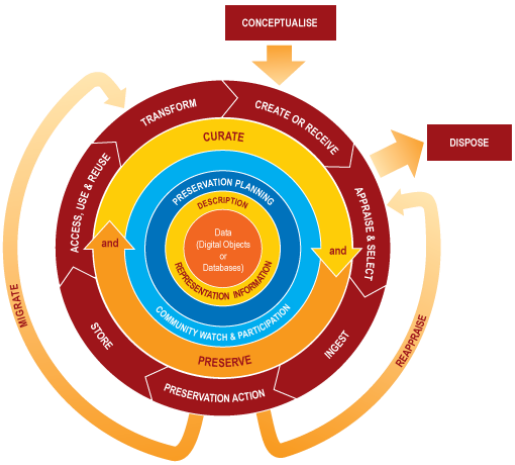
Data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyse or add to the data, and data may be re-used by other researchers. Well organised, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation.



1.2.3 Data Curation Centre Curation Lifecycle

The DCC Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle. You can use the model to plan activities within your organisation or consortium to ensure that all of the necessary steps in the curation lifecycle are covered.

It is important to note that the model is an ideal. In reality, users of the model may enter at any stage of the lifecycle depending on their current area of need. For example, a digital repository manager may engage with the model for this first time when considering curation from the point of ingest. The repository manger may then work backwards to refine the support they offer during the conceptualisation and creation processes to improve data management and longer-term curation.



§2 Data Management Systems

There are myriad of solutions that could be used to manage your research data. The choice you make will vary depending on the environment where you will collect your data. For example, if you are conducting a household survey then you may be best to utilise a mHealth solution. However, if you have a lengthy survey that you would like to be able to jump between sections then you may choose to use paper with Optical Character Recognition (OCR) data entry. The support offered within each solution also varies, for example OpenClinica offers full lifecycle management of research data whereas other solutions require you to develop bespoke bolt-ons to replicate the missing functionality such as managing data clarifications (cleaning) and capturing audit trails. These are solutions linked to the data collection method:

2.1 Systems for collecting and managing data:

2.1.1 Paper based with manual data entry

This method of data collection can include a variety of systems to manage your data. Typically this would include a relational database management system with data entry screens. You may include double-data entry with a third person verifying the correct value where differences have been entered. Crucially this does not capture when the error is entered incorrectly by both data entry staff. Solutions currently include:

Solution	License Cost involved?	Database	Hosting	Validation status	DM Lifecylce covered	LSTM Contact for more information
MACRO	Yes	MS SQL	In-house or can be cloud-based	Out of the box	Full	James Smedley
Open Clinica	Yes - for validated enterprise version No - for open source solution	MySQL / PostgreSQL		Out of the box for enterprise. For open source, testing and validation would need to be devised in conjunction with the development of the system itself.	Full	James Smedley
EpiInfo	No – Open source	MS Access	Desktop application			Caroline Jeffery
MS Access	Yes only for developer, can be deployed	MS Access / MS SQL / MySQL	Desktop application but can link to ODBC	Testing and validation would need to be devised in conjunction with the	Partial – needs development of cleaning &	James Smedley

	using runtime environment		compliant database such as MS SQL server.	development of the system itself.	coding to manage data	

2.1.2 Paper based with OCR data entry

This method of data management utilises optical character recognition (OCR). Typically this would include the creation of a paper questionnaire using a form designer. This template would then be distributed for data collection. Upon completion the data entry staff would then scan the completed form and the system would then evaluate the form and ask the user to make changes based on system defined confidence levels. Solutions currently used at LSTM:

Solution	License Cost involved	DB	Hosting	Validation status	DM Lifecylce covered	LSTM Contact for more information
Formic	Yes	Excel	Hosted at LSTM	Testing and validation would need to be devised in conjunction with the development of the system itself.	Data entry	Barbara Madaj
TeleForm	Yes	MS SQL	Hosted at MLW and Eijkman Institute of Molecuar Biology, Indonesia.	Testing and validation would need to be devised in conjunction with the development of the system itself.	Data entry	James Smedley

2.1.3 Electronic Data Capture / eHealth / mHealth

Electronic data capture systems can increase the accuracy and completeness of your data as skip patterns and range checks are incorporated at the point of capture. The importance of pre-testing and validating the survey as fit-for-purpose is a key milestone here as an incorrect skip pattern or range check could lead to entire missed sections or invalid data.

There are a range of solutions currently deployed at LSTM:

Solution	License Cost involved	OS & DB	Hosting	Validation status	DM Lifecylce covered	LSTM Contact for more information
Open Data Kit (ODK)	No	Android / MySQL or Postgres	In-house or can be cloud-based	Testing and validation would need to be devised in conjunction with the	Data entry	James Smedley

				development of the system itself.		
Enketo	No	Windows / MySQL or Postgres	In-house or can be cloud-based	Testing and validation would need to be devised in conjunction with the development of the system itself.	Data entry	James Smedley
Magpi	Yes	Android / MySQL	cloud-based	Testing and validation would need to be devised in conjunction with the development of the system itself.	Data entry	James Smedley
FileMaker Pro	Yes	Apple / MS SQL	cloud-based	Testing and validation would need to be devised in conjunction with the development of the system itself.	Data entry	Barbara Madaj

2.1.3.1 Smart phones vs Dumb phones

Whilst the explosion of smart phones has seen a rapid rise in the availability of apps and survey tools, the use of 'dumb' phones for electronic data collection remains a possibility, especially in LMIC setting. These include systems such as:

- Frontline SMS
- RapidPro

§3 Information Security

3.1 Securing your data

It is important that all staff take due care to safeguard the security of LSTM research data as they are valuable assets that you will invest considerable time, effort and money in creating during the lifetime of the research project. Protecting your data from loss is therefore an important aspect of research data management and careful consideration should be taken in which environment you choose to store your data, what type of backup and retention systems to implement and levels of access that should be granted. This applies at all stages within the lifecycle of the project.

3.2 LSTM Information Classification Matrix

We have devised an information classification matrix (see Appendix A) to assist you in establishing a framework for classifying your data based on its level of sensitivity, value and criticality to LSTM as required by the Code of Practicee for the Acceptable Use of Computer & IT Facilities at LSTM. Classification of information will aid in determining baseline security controls for the protection of data. For example, in helping to decide whether to transfer a file electronically via email or an encrypted email (see Appendix B).

3.3 Recommendation for documentation

All LSTM staff have the facility to use One Drive for Business and we recommend that you always save important documents to this area of your hard drive. This area of your hard drive is automatically backed up and all staff have a minimum of 25GBs of file storage and this can be extended if required. Additionally, any files that you save in this area can be accessed remotely from any internet connected computer by going to http://portal.lstmed.ac.uk

3.4 Sharing documents internally

One Drive for Business is designed to store the files that are personal to you and not for files that are shared within a group. If you have files that you would like to share with colleagues internally then utilising a SharePoint site would be the better option. If you do not have a team site or if you want to learn about good examples of team-sites utilised at LSTM please contact Tom Cowling (tom.cowling@lstmed.ac.uk) in LSTM IT Services.

The documents shared this way can also be synced to your hard drive so that when you're travelling without an internet connection you still have access to them please see user guide for more information.

3.5 Sharing documents externally

Whilst you can create external accounts for your SharePoint team-site, the syncing of files to the hard drive is not available for external accounts. External members would be required to login to the team-site to view documents.

There are other tools available for sharing documents / data depending on their data classification / sensitivity, such as Dropbox or Sugarsync. If you are in doubt about whether the system you are

planning to use meets minimum security requirements then please contact the Director of IT Services (Eric Healing).

3.6 Other storage options

LSTM do not recommend the use of removal media for file storage. Please speak to LSTM IT Services before using this form of storage as it is likely that they can provide a more secure method of storage. It is also not advised that you store sensitive, personal data on a portable drive or on a laptop. Your attention is drawn to your responsibilities under the Data Protection Act, 1998. Further information and support on data protection can be found here and on the Information Commissioner's website. See also the new GDPR responsibilities that came into force as of 25 May 2018: https://gdpr-info.eu/

3.7 Recommendation for patient identifiable data

Data that would allow the identification of individuals should not be stored on mobile device of any kind, this includes laptops. For sensitive data like this it should be kept in the country of origin and anonymised prior to transferring back to LSTM.

In the event of a non-anonymised data set being transferred to LSTM this should not be stored on the One Drive for business storage please contact LSTM IT Services for more information.

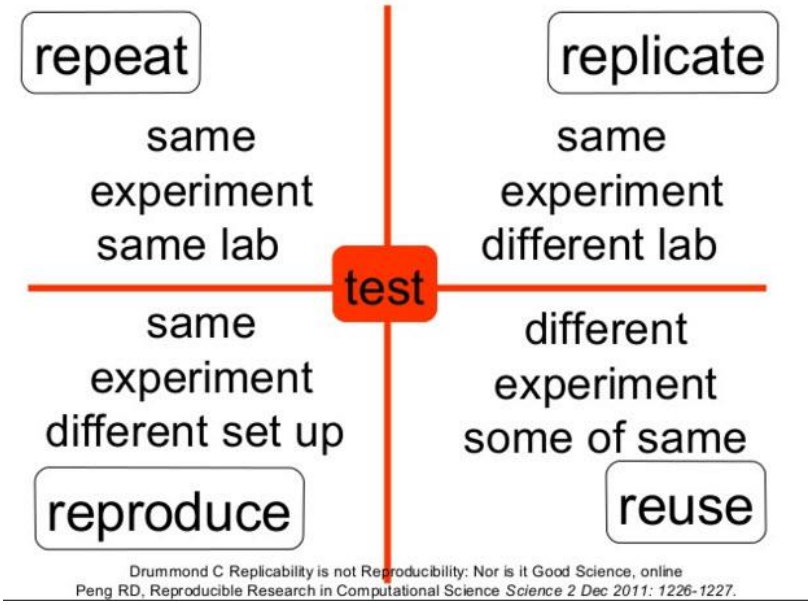
3.8 Anonymising your data

Anonymisation is the process of turning data into a form which does not identify individuals and where identification is not likely to be able take place. A data set is not considered anonymised if you can still identify a participant from it. Typically you would remove names, initials, addresses and date of births and would then consider the data anonymised. However, the remaining information may be so specific to one individual which would then not be considered anonymised. Some further guidance is available in LSTM's Guide to anonymisation in clinical studies (see Appendix C).

§4 Archiving and Sharing

With the development of meta-analysis and funder requirements for sharing data, research data remains a valuable resource, even after the original study it was generated from comes to a close. This enables future researchers to open up new lines of enquiry without the duplication of effort involved in regenerating the data. Metadata plays a key role in allowing other researchers and you yourself when looking back in the future, to make sense of your data.

4.1 Benefits & drivers for managing and sharing your data



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

By managing your data well, you can, repeat, replicate, reproduce and reuse your data and:

- Find and understand it when needed now and in the future.
- Avoid unnecessary duplication.
- Validate your results when required.
- Ensure your research is visible and has impact.
- Get credit when others cite your work.
- Comply with Funder mandates.
- Research Councils UK in its 2009 code of good research conduct says data should be preserved and accessible for 10 years +Research Funder data policies increasingly demanding of institutional commitment and encourage or mandate the creation of a research data management plan and the deposit of research data in a recognised data centre where such exist.
- Many leading journals require underlying datasets also to be published or made accessible as part of the essential evidence base of a scholarly article.

4.2 LSTM Data Repository

LSTM will have its own data repository later this year. The data repository will use the Eprints software making it compliant with funders' open access policies through use of the internationally recognised OAI 2.0 standard. It will provide a central registry of LSTM's published data which will enable us to locate, manage and curate the data after the project has ended.

It is planned that the data repository will have the following features:

- Facility to mint a DOI via Datacite
- Links to the publications repository (Online Archive)
- Archival storage to meet your funder's requirements

If you have any data sets which require archiving, or are submitting a new grant applications and need to include costs, please contact julia.martin@lstmed.ac.uk to discuss your requirements.

4.3 External archives

After checking your funders requirements, such as Open Access Initiative (OAI) compliance for example, the Register of research data repositories (r3data) has developed an excellent tool for you to identify a suitable repository for your data outputs. If you choose to submit your data to an external repository, LSTM would require you to publish a basic entry in our research data archive to ensure we have a record for adminstrative purposes.

Appendix

A: LSTM Information Classification Matrix

LSTM Information Classification Matrix

What is the Information Classification Matrix?

It is a document designed to help you ensure that you understand the difference between information that can be shared publically, internally or must remain confidential or restricted.

How to use the Information Classification Matrix

1 Identify which column covers the information / data you wish to handle.

2 You do this by looking at the Key at the top of the Classification Matrix

For example, let's say that you need to send the hard copy personnel file of a staff member to another person in the LSTM. You would look at the key in the Classification Matrix and see that as you are dealing with person-identifiable information you need to be looking in the "Confidential" column

3 Now you need to look down the first and second columns of the Classification Matrix until you find the action that you want to take. In the example we are using this would be "Transmission by Post, Fax or e-mail", "Mail within the LSTM (i.e. between buildings)"

4 Now you need to look along the row until you get to the relevant column which lists what you must do.

In the example we are using this would be the "Confidential" column where you would find that you must send the personnel file in a "Sealed inter-office envelope marked Confidential".

Definitions

Public	'Public' information can be disclosed or disseminated without any restrictions on content, audience or time of publication. Disclosure or dissemination of the information must not violate any applicable laws or regulations, such as privacy rules. Modification must be restricted to individuals who have been explicitly approved by information owners to modify that information, and who have successfully authenticated themselves to the appropriate computer system.
Internal	'Internal use' information can be disclosed or disseminated by its owner to appropriate members of LSTM, partners and other individuals, as appropriate by information owners without any restrictions on content or time of publication.
Confidential	'Confidential' information has significant value for LSTM, and unauthorized disclosure or dissemination could result in severe financial or reputational damage to the School, including fines of up to £500,000 from the Information Commissioner's Office, the revocation of research contracts and the failure to win future research bids. Data that is defined by the Data Protection Act as Sensitive Personal Data falls into this category. Only those who need explicitly need access must be granted it, and only to the least degree in order to do their work (the 'need to know' and 'least privilege' principles). When held outside

	LSTM, on mobile devices such as laptops, tablets or phones, or in transit, 'Confidential' information must be protected. Specific levels of protection are indicated in the matrix below.
Restricted	'Restricted' information requires the same protection as 'Confidential' information, but in addition, it is subject to controls on access, such as only allowing valid logons from a small group of staff. 'Restricted' information must be held in such a manner that prevents unauthorised access i.e. on a system that requires a valid and appropriate user to log in before access is granted. Unauthorised disclosure is likely to result in significant adverse impact, embarrassment or penalties to LSTM, its stakeholders or employees.

Further examples

Example 1: Medical information

You have an e-mail regarding the medical information about one of your staff members. As this is classed as "Sensitive" data under the Data Protection Act, 1998 it must be treated as confidential. The example below gives guidance to how this data should be handled:

2. Transmission by post, messaging service or e-mail

	Public	Internal	Confidential	Restricted
Internal mail within LSTM.	No special handling required.	No special handling required.	Sealed internal envelope marked "Private & Confidential"	Sealed internal envelope marked "Restricted and Confidential". Notify recipient in advance.
Mail outside of LSTM	No special handling required.	2nd class mail. No special handling required.	2nd class mail. Marked "Private and Confidential" with return address on the envelope. Traceable delivery is preferred e.g. Recorded or Special delivery.	Marked "Private and Confidential". Traceable delivery preferred, e.g. Recorded or Special delivery.
E-mail within LSTM	No special handling required.	No special handling required.	Refrain from use of personal data. Use of e-mail discouraged where practical.	Use of any personal data is prohibited. Use of e-mail strongly discouraged, unless encrypted.
E-mail outside of LSTM	No special handling required.	No special handling required.	Use of e-mail containing personal data prohibited unless encrypted or emergency situation. Use of e-mail strongly discouraged. Broadcast to distribution lists is prohibited.	Use of e-mail containing personal data prohibited unless encrypted or emergency situation. Use of e-mail strongly discouraged. Broadcast to distribution lists is prohibited.

Example 2: Information subject to a Confidentiality agreement

This could cover an arrangement with an external company or research partner to enable either party to share proprietary information for potential collaborations. If this information (data, material, tools, drawings, etc.) is subject to a confidentiality agreement, then this will mean that you must follow the guidance in the "Confidential" column of the matrix. The confidentiality agreement itself will contain specific safeguards. Specific examples would include:

- All electronic documents should be stored on secure media & should not be stored on shared drives (i.e. S:/ drive) without password protection;
- Photocopying is strongly discouraged and should be carried out only when necessary;
- If carrying out a Skype for Business call (or similar), you should avoid being overheard. If you normally work in an open office environment then you should select a secure area for the call.
- If sharing proprietary information in a face-to-face meeting, you should follow up, in writing, that the information shared is confidential.

B: Encryption

Data Security Advice

It is essential that staff ensure all confidential data held on Computers, Laptops or any removable media e.g. external hard disks, USB/Pen Drives and DVD/CD is secured using encryption software and password protection.

- All Computers, Laptops should have an encrypted hard drive, the standard LSTM laptops for all laptop issued from 31 March 2015 have encrypted hard drives. If you laptop was not issued from IT Services or was issued before this date IT Services can encrypt your hard drive for you.

- USB devices should be encrypted.

If you require any further advice or support please submit a Support Request

Encrypted Data Pens

Staff and students should make use of AES encrypted pen drives to ensure the security and integrity of valuable data. IT Services recommends the Integral Crypto Drive, which features industry leading security features, and an easy-to-use software interface. These are designed for PC users, MAC versions are available on request but cannot be used cross-platform.

The Crypto drive is available to purchase from Insight (supplier code INS08) using an ISF. Alternatively you can place your order with the Purchasing Department through CODA.

The Crypto Drive specifications are:

- **Military Level Security** – AES 256 bit hardware encryption. Mandatory encryption of all files (100% privacy)
- **Secure Entry** - Data cannot be accessed or removed without the correct high strength 6-16 character password
- **Brute-force Password Attack Protection** – Data will be automatically erased after 6 failed access attempts and Drive reset
- **Personal ID Function (optional)** – Contact details can be added so that Drive can be returned, whilst confidential data remains secure
- **Zero Footprint** – No software installation required
- **Rugged Design** - Steel inner casing and rubberised outer casing, designed to protect the data if dropped, crushed, splashed, or subjected to tampering attempts
- **Easy To Use** - Pre-loaded user interface in 22 languages

C: Guide to Anonymisation in Clinical Studies

What is anonymisation?

Data in clinical studies should be modified so that it is not possible for individuals to be identified, this is known as anonymisation.

Why should data be anonymised?

GCP

Section 2.11 ICH GCP, the confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).

Law

The Data Protection Act would apply if the data contained personal information and its principles must be adhered to.

Ethical Approval

Any access to personal data will be scrutinised by the Research Ethics Committee and the use of anonymised data will be much easier to justify both ethically and legally.

When should data be anonymised?

In general, data should be anonymised as soon as it is feasible. Although anonymisation may induce delays and increase the risks of error, even a simple coding system provides a safeguard against accidental or mischievous release of confidential information.

However, it is most important in situations where there the data is of a sensitive nature (e.g. HIV status), and the release of personal information could have a negative effect on its owner.

High risk situations include:

When data is transferred or shared with other researchers

Where research data are shared with other researchers for new studies, the custodian must ensure that the group accepts a duty of confidence and protects confidentiality.

When data is transported overseas

Where identifiable data is passed on, the custodian must ensure that these are not passed to a country without legal protection for identifiable personal data equivalent to that in the UK, unless the custodian first assures themselves that the data will be adequately protected in practice. Under the terms of the Data Protection Act 1998, there are no special restrictions on transfers of identifiable data within the European Economic Area (EEA).

Outside of the EEA, e.g. for data sent to the USA, or to a developing country, the custodian must either:

- Anonymise the data
- Obtain the individual explicit consent to send their data to another centre
- Remain able to control the use of the data transferred, however, this is not always possible to ensure, owing to differing legal frameworks and the need for monitoring.

When unauthorised access might adversely affect the welfare of an individual

This could include study data involving details of someone's sexual health or any politically and culturally sensitive information.

How should data be anonymised?

Remove direct identifiers

Anything that could have the reasonable potential to be used to identify individuals must be removed. Personal Information could contain any of the following;

- name,
- address,
- phone number,
- email address,
- postcode,
- any identifying reference number (e.g. NHS number), photograph, names of relatives.

geo-referenced data – spatial references (point coordinates, small areas) may disclose the position of individuals.

However there are lots of other potential identifiers, such as;

- rare disease or treatment, especially if an easily noticed illness/disability is involved,
- partial post-code/address,
- rare occupation or place of work,
- combinations of birth date, ethnicity, place of birth and date of death, unique or exceptional values (outliers) or characteristics.

Tips for anonymising data:

Reduce the precision/detail of a variable through aggregation

Examples:

birth year versus date of birth, occupational categories, area rather than village.

Generalise meaning of detailed text variable

Example: occupational expertise

Restrict upper and lower ranges of a variable to hide outliers

Example: income and age

Combine variables:

Example: consider creating non-disclosive rural/urban variable from place variable.

Reduce the precision of spatial references

Replace point coordinate with larger, non-disclosing geographical areas, e.g. km^2 area, postcode district, ward, road.

Reduce the precision of point coordinates with meaningful variable typifying the geographical position, e.g. catchment area, poverty index or population density.

Researchers must also take into account the sample size, the ways in which the results will be published, the number of staff who will have access and any other circumstances that might lead to identification of individuals.

To what degree should data be anonymised?

There are various levels of anonymisation:

Coded Information

This is where the team working with the information has coded any identifiable data, but they also have access to the key to uncode this information. This helps to meet legal and ethical obligations but still falls within the scope of the Data Protection Act.

Linked Anonymised Data

This is where the data is anonymous to the research team that holds it, but contains coded information which could be used to identify people. The key to the code, might, for example be held by those responsible for the individual's care or by the custodians of a larger research database or register.

Unlinked Anonymised Data

Contains nothing that has reasonable potential to be used to identify individuals and any links have been irreversibly broken.

The degree of Anonymisation required will have to be assessed by the PI depending on the specific circumstances. For example:

In small scale clinical trials, which involve frequent reference by research staff to current patients' conditions, encoding and decoding information can present a significant obstacle to effective team work, and increases the risk of error that could affect the patient's care. Use of weaker codes (such as initials) in processing research data is acceptable where patients have already consented to the use of their information in research as well as for their care, and when it can be guaranteed that only a small number of research staff will have access to the information.

Top Tips

- never disclose personal data, unless there is consent for disclosure – always consider anonymisation of research data together with consent agreements and access restrictions,
- regulating/restricting user access may offer a better solution than anonymising,
- avoid collecting data that need anonymisation e.g. don't ask for full names if they later need to be removed from data
- remove, mask, change direct identifiers,
- maintain maximum meaningful information,
- where possible, replace data with code rather than remove,
- re-users of data have the same legal and ethical obligations to NOT disclose confidential information as the primary users.
- retain unedited versions of data for preservation,
- plan anonymising at the start of the research, not at the end.

Further Information

Please contact the Research Office.

Natalie Strickland, email: nswann@liv.ac.uk or Sian Roberts s.roberts@liv.ac.uk